न्याय दर्शन से AI नैतिकता: प्राचीन तर्कशास्त्र, आधुनिक समाधान

(Justice Philosophy to AI Ethics: Ancient Logic, Modern Solutions)



लेखक: *Anand Singh Dhami*

प्रस्तुतकर्ता: *Indian Philosophy & Ethics*

विषय क्षेत्र: न्याय दर्शन, कृत्रिम बुद्धिमत्ता नैतिकता, प्रमाण सिद्धांत, पंचावयवी न्याय, एल्गोरिदमिक न्याय, दार्शनिक AI, भारतीय दर्शन एवं प्रौद्योगिकी, AI पारदर्शिता, जवाबदेही

संक्षिप्त परिचय:- *यह शोध पत्र कृत्रिम बुद्धिमत्ता (AI) की नैतिक चुनौतियों-जैसे डेटा पक्षपात, अपारदर्शिता और जवाबदेही के संकट के समाधान हेतु **न्याय दर्शन** के तार्किक एवं ज्ञान-मीमांसात्मक ढाँचे का विश्लेषण करता है। पश्चिमी नैतिक सिद्धांतों की सीमाओं को पहचानते हुए, यह अध्ययन न्याय के **प्रमाण सिद्धांत, पंचावयवी न्याय** और **हेत्वाभास** को AI की तकनीकी विफलताओं के दार्शनिक विकल्प के रूप में प्रस्तुत करता है।*
*शोध के निष्कर्ष दर्शाते हैं कि पक्षपाती डेटा 'अशुद्ध शब्द प्रमाण' है और AI की 'ब्लैक बॉक्स' समस्या 'हेतु' एवं 'उदाहरण' की अनुपस्थिति का परिणाम है। इन अंतर्दृष्टियों के आधार पर, यह पत्र एक **"न्याय-आधारित AI ऑडिट एवं डिजाइन ढाँचा"** प्रस्तावित करता है, जो AI जीवनचक्र में पारदर्शिता और न्यायसंगतता सुनिश्चित करने के लिए एक मौलिक भारतीय दार्शनिक दृष्टिकोण प्रदान करता है।*

🔗 ***ऑनलाइन लेख: -*** *https://indianphilosophyethics.blogspot.com/2026/01/nyaya-darshan-ai-ethics.html*

न्याय दर्शन से AI नैतिकता: प्राचीन तर्कशास्त्र, आधुनिक समाधान

1. परिचय

कृत्रिम बुद्धिमत्ता (AI) का तीव्र विकास 21वीं सदी की सबसे परिभाषित तकनीकी उन्नतियों में से एक है। चिकित्सा निदान और स्वायत्त परिवहन से लेकर सामग्री निर्माण एवं न्यायिक सहायता तक, AI प्रणालियाँ मानव जीवन के प्रत्येक क्षेत्र में गहराई से समा चुकी हैं। हालाँकि, इस अभूतपूर्व प्रगति के साथ गंभीर नैतिक चुनौतियाँ भी उत्पन्न हुई हैं। एल्गोरिदमिक पक्षपात, गोपनीयता का क्षरण, पारदर्शिता का अभाव ("ब्लैक बॉक्स" समस्या), तथा निर्णयन प्रक्रिया में मानवीय नियंत्रण व जवाबदेही का संकट आज AI अनुसंधान एवं नीति के केंद्र में हैं । इन चुनौतियों के समाधान हेतु प्रस्तावित अधिकांश नैतिक ढाँचे (जैसे निष्पक्षता, जवाबदेही, पारदर्शिता के सिद्धांत) मुख्यतः पश्चिमी दार्शनिक परंपरा विशेष रूप से उपयोगितावाद (Utilitarianism), कर्तव्य-नैतिकता (Deontology), तथा न्याय के सिद्धांतों (Rawls, 1971) से प्रेरित हैं।

किंतु, AI एक वैश्विक तकनीक है जिसका प्रभाव विविध सांस्कृतिक एवं दार्शनिक परिवेशों में पड़ता है। एक समावेशी एवं सार्थक वैश्विक AI नैतिकता के लिए विभिन्न दार्शनिक परंपराओं के ज्ञान को सम्मिलित करना अनिवार्य है। वर्तमान में, इस दिशा में एक महत्वपूर्ण अंतराल विद्यमान है। जैसा कि हाल के शोध से पता चलता है, "गैर-पश्चिमी ज्ञान-मीमांसाएँ (epistemologies) अभी भी इन बहसों से बहिष्कृत हैं" । यह बहिष्कार न केवल एक बौद्धिक कमी है, बल्कि इससे तकनीकी समाधानों की प्रभावकारिता भी सीमित होती है, क्योंकि पश्चिम-केंद्रित दृष्टिकोण सार्वभौमिक नैतिक समस्याओं के सभी पहलुओं को समझने में असमर्थ हो सकते हैं।

इसी संदर्भ में, यह शोध पत्र भारतीय दर्शन की न्याय परंपरा पर ध्यान केंद्रित करता है, जो षड्दर्शनों में से एक है। प्लेटो या अरस्तू के तर्कशास्त्र के समानांतर, न्याय दर्शन का प्राथमिक उद्देश्य वैध ज्ञान (प्रमा) प्राप्त करने के साधनों एवं तार्किक नियमों का विवेचन करना है। यह दर्शन अमूर्त आध्यात्मिक विमर्श से आगे बढ़कर, प्रमाण (ज्ञान के स्रोत), अनुमान (तर्क), और तर्क-दोषों (हेत्वाभास) की पहचान पर एक संरचित, विश्लेषणात्मक प्रणाली प्रदान करता है।

इस शोध पत्र का मूल अनुसंधान प्रश्न यह है: क्या न्याय दर्शन द्वारा प्रतिपादित ज्ञान-मीमांसात्मक (epistemological) एवं तार्किक ढाँचा, समकालीन AI प्रणालियों में निहित नैतिक चुनौतियों के विश्लेषण एवं समाधान हेतु एक व्यवहार्य दार्शनिक प्रतिमान प्रस्तुत कर सकता है? इस प्रश्न का उत्तर देने के लिए, हम निम्नलिखित उद्देश्यों पर कार्य करेंगे:

1. न्याय दर्शन की प्रमुख अवधारणाओं (प्रमाण, पंचावयवी, हेत्वाभास) का एक स्पष्ट विवेचन प्रस्तुत करना।
2. इन अवधारणाओं का AI प्रणालियों की तकनीकी प्रक्रियाओं (डेटा प्रोसेसिंग, मॉडल ट्रेनिंग, अनुमान) तथा उनसे उत्पन्न नैतिक समस्याओं के साथ तुलनात्मक विश्लेषण करना।
3. उपरोक्त विश्लेषण के आधार पर, AI के डिजाइन, विकास, परीक्षण एवं नियमन प्रक्रियाओं के लिए एक "न्याय-आधारित AI ऑडिट एवं डिजाइन ढाँचा" प्रस्तावित करना।

इस प्रकार, यह अध्ययन एक अंतःविषयक (Interdisciplinary) दृष्टिकोण अपनाता है जो प्राचीन भारतीय चिंतन एवं अत्याधुनिक प्रौद्योगिकी के बीच एक सार्थक संवाद स्थापित करने का प्रयास करता है। हमारी मूल परिकल्पना यह है कि न्याय दर्शन का तार्किक कठोरता पर आधारित ढाँचा, AI की नैतिक समस्याओं को उनकी जड़ यानि दोषपूर्ण ज्ञान-प्रक्रियाओं तक पहुँचकर समझने और सुलझाने में सक्षम है।

2. साहित्य समीक्षा

AI नैतिकता पर साहित्य का तेजी से विस्तार हो रहा है, जिसे मोटे तौर पर तीन व्यापक श्रेणियों में वर्गीकृत किया जा सकता है: (क) तकनीकी-केंद्रित दृष्टिकोण, (ख) नियामक एवं नीतिगत दृष्टिकोण, तथा (ग) दार्शनिक एवं सामाजिक-सांस्कृतिक दृष्टिकोण।

तकनीकी-केंद्रित दृष्टिकोण मशीन लर्निंग एल्गोरिदम में निष्पक्षता (फेयरनेस) को मापने और बढ़ाने, व्याख्यायोग्य AI (XAI) विकसित करने, और डेटा गोपनीयता (जैसे डिफरेंशियल प्राइवेसी) सुनिश्चित करने के उपायों पर केंद्रित है । जबकि ये उपाय आवश्यक हैं, ये अक्सर लक्षणों का इलाज करते हैं न कि रोग का; ये पद्धतियाँ उन गहरे दार्शनिक प्रश्नों जैसे "न्याय क्या है?" या "वैध ज्ञान कैसे प्राप्त होता है?" को संबोधित नहीं करतीं, जिन पर AI की नैतिकता आधारित होनी चाहिए।

नियामक एवं नीतिगत दृष्टिकोण - यूरोपीय संघ का AI एक्ट, UNESCO की AI नैतिकता पर सिफारिशें, और विभिन्न राष्ट्रीय AI रणनीतियों जैसे ढाँचे बनाने पर केंद्रित है। ये दस्तावेज़ उच्च-स्तरीय सिद्धांत (जैसे मानव निगरानी, सामाजिक कल्याण) प्रस्तुत करते हैं, लेकिन अक्सर इन सिद्धांतों को व्यावहारिक तकनीकी मानकों में अनुवाद करने की चुनौती बनी रहती है।

दार्शनिक एवं सामाजिक-सांस्कृतिक दृष्टिकोण- AI के नैतिक आधारों की खोज करता है। प्रमुख पश्चिमी नैतिक सिद्धांत उपयोगितावाद (सर्वाधिक लोगों का अधिकतम कल्याण), कर्तव्यनैतिकता (सार्वभौमिक नियमों का पालन), तथा सद्गुण नैतिकता (चरित्र निर्माण) का AI संदर्भ में व्यापक विश्लेषण किया गया है । हाल के वर्षों में, गैर-पश्चिमी परंपराओं के योगदान पर भी ध्यान दिया जाने लगा है। उदाहरण के लिए, स्वदेशी नैतिकता के ढाँचे, जो पारस्परिकता (reciprocity), सामुदायिकता और प्रकृति के साथ संतुलन पर जोर देते हैं, को AI नैतिकता के लिए एक महत्वपूर्ण प्रतिकारक के रूप में प्रस्तावित किया गया है । इसी तरह, हिंदू दार्शनिक अवधारणाओं जैसे अहिंसा (करुणा), लोकसंग्रह (सार्वजनिक कल्याण), और अद्वैत (अंतर्निर्भरता) का AI के संदर्भ में अन्वेषण शुरू हुआ है ।

अंतराल: इन प्रगतियों के बावजूद, एक स्पष्ट अंतराल है। न्याय दर्शन, जो स्पष्ट रूप से तर्क, ज्ञान-मीमांसा और विश्लेषणात्मक विधि पर केंद्रित है, का AI नैतिकता के साथ व्यवस्थित एवं गहन समीक्षा का अभाव है। जहाँ वेदांत या बौद्ध दर्शन पर कुछ साहित्य मौजूद है, वहीं न्याय दर्शन के ग्रंथों (गौतम के न्याय-सूत्र, वात्स्यायन के न्याय-भाष्य) में निहित औपचारिक तार्किक प्रणाली को AI की समस्याओं पर लागू करने का प्रयास सीमित है। यह शोध पत्र इसी महत्वपूर्ण अंतराल को भरने का प्रयास करता है, यह दर्शाते हुए कि न्याय दर्शन न केवल एक नैतिक दृष्टिकोण, बल्कि एक व्यावहारिक विश्लेषणात्मक उपकरण भी प्रदान करता है जो AI सिस्टम की अंतर्निहित तार्किक प्रक्रियाओं को सीधे संबोधित कर सकता है।

3. शोध पद्धति

यह शोध एक गुणात्मक (Qualitative) तथा सैद्धांतिक (Theoretical) अनुसंधान है, जिसका उद्देश्य एक नए अवधारणात्मक ढाँचे का निर्माण करना है। इसमें निम्नलिखित पद्धतिगत चरण अपनाए गए हैं:

1. **वैचारिक विश्लेषण (Conceptual Analysis)**: प्राथमिक दार्शनिक ग्रंथों (मुख्यतः गौतम के न्याय-सूत्र एवं वात्स्यायन के न्याय-भाष्य) के आधार पर न्याय दर्शन की प्रमुख अवधारणाओं प्रमाण, पंचावयवी, हेत्वाभास का स्पष्टीकरण एवं विवेचन।

2. **तुलनात्मक अध्ययन (Comparative Study)**: उपरोक्त अवधारणाओं का समकालीन AI प्रणालियों की तकनीकी प्रक्रियाओं तथा प्रकाशित नैतिक विफलताओं के केस-स्टडीज के साथ तुलनात्मक विश्लेषण। यह तुलना सारणीबद्ध रूप में प्रस्तुत की गई है (तालिका 1 देखें)।

3. **अंतःविषयक संश्लेषण (Interdisciplinary Synthesis)**: दर्शनशास्त्र, कम्प्यूटर विज्ञान एवं नीति विज्ञान के सिद्धांतों को एकीकृत करते हुए एक नए प्रतिमान (Framework) का निर्माण। यह प्रतिमान AI जीवनचक्र के विभिन्न चरणों के लिए व्यावहारिक दिशा-निर्देश प्रदान करता है।

आँकड़ों के स्रोत: यह शोध प्राथमिक रूप से दो प्रकार के स्रोतों पर निर्भर करता है:

प्राथमिक स्रोत: न्याय दर्शन के मूल संस्कृत ग्रंथ एवं उनके मान्यता प्राप्त भाष्य।

द्वितीयक स्रोत: (क) AI नैतिकता, एक्सप्लेनेबल AI (XAI), और एल्गोरिदमिक न्याय पर समकालीन शैक्षणिक साहित्य। (ख) AI सिस्टम्स की नैतिक विफलताओं के प्रकाशित केस-स्टडीज (जैसे, भर्ती एल्गोरिदम में लैंगिक पक्षपात, स्वास्थ्य देखभाल एल्गोरिदम में नस्लीय पक्षपात)।

इस पद्धति का चयन इसलिए किया गया है क्योंकि शोध का उद्देश्य मात्रात्मक डेटा का विश्लेषण करना नहीं, बल्कि अवधारणाओं के बीच नए संबंध स्थापित करते हुए एक सैद्धांतिक ढाँचे का प्रस्ताव करना है।

4. न्याय दर्शन का दृष्टिकोण: प्रासंगिक अवधारणाओं का विश्लेषण

AI नैतिकता के विश्लेषण हेतु न्याय दर्शन की तीन मूलभूत अवधारणाएँ विशेष रूप से प्रासंगिक हैं। इनका संक्षिप्त विवेचन यहाँ प्रस्तुत है:

4.1 प्रमाण सिद्धांत : वैध ज्ञान के स्रोत

न्याय दर्शन के अनुसार, सही निष्कर्ष (न्याय) केवल तभी संभव है जब प्रारंभिक ज्ञान वैध हो। यह वैध ज्ञान चार प्रमाणों से प्राप्त होता है:

प्रत्यक्ष (प्रत्यक्ष ज्ञान): इंद्रियों के माध्यम से वस्तु के सीधे संपर्क से उत्पन्न ज्ञान। यह आधारभूत है।

अनुमान (अनुमान): एक ज्ञात साक्ष्य (हेतु/लिंग) के आधार पर एक अदृश्य लक्ष्य (साध्य) के बारे में निष्कर्ष निकालना। (उदाहरण: पहाड़ी पर धुआँ देखकर आग का अनुमान)।

उपमान (तुलना): किसी अज्ञात वस्तु की किसी ज्ञात वस्तु से समानता के आधार पर पहचान करना।

शब्द (विश्वसनीय स्रोत): एक विश्वसनीय साक्षी (आप्त) या प्रामाणिक ग्रंथ से प्राप्त ज्ञान।

4.2 पंचावयवी न्याय (Five-Membered Syllogism): संरचित तर्क की पद्धति

किसी भी तर्क या दावे की वैधता सिद्ध करने के लिए, न्याय दर्शन इसे पाँच क्रमिक एवं अन्योन्याश्रित चरणों में प्रस्तुत करने की विधि सिखाता है:

1. **प्रतिज्ञा**: वह कथन जिसे सिद्ध करना है। (थीसिस)

2. **हेतु**: उस कथन का कारण या प्रमाण देना। (ग्राउंड/रीज़न)

3. **उदाहरण**: एक सार्वभौमिक नियम प्रस्तुत करना और एक ज्ञात उदाहरण देना जो उस नियम का अनुसरण करता हो। (यूनिवर्सल प्रोपोज़िशन एवं उदाहरण)

4. **उपनय**: वर्तमान मामले को उस सार्वभौमिक नियम से जोड़ना, यह दिखाते हुए कि यहाँ भी वही संबंध लागू होता है। (अप्लिकेशन)

5. **निगमन**: पहले के चरणों के आधार पर प्रतिज्ञा की पुष्टि करते हुए निष्कर्ष निकालना। (कंक्लूज़न)

यह संरचना तर्क को पारदर्शी, परखने योग्य और विवाद के लिए खुला बनाती है।

4.3 हेत्वाभास (Logical Fallacies): तर्क दोषों का वर्गीकरण

न्याय दर्शन केवल सही तर्क ही नहीं, बल्कि गलत तर्कों को पहचानने के नियम भी देता है। यह अनुमान में होने वाले दोषों (हेत्वाभास) का विस्तृत वर्गीकरण करता है, जैसे:

सव्यभिचार हेतु: ऐसा कारण जो स्थिर न हो, जो कभी साध्य के साथ तो कभी उसके बिना भी पाया जाए। (अनियमित/असंगत कारण)।

विरुद्ध हेतु: ऐसा कारण जो सिद्ध करने योग्य वस्तु के विपरीत सिद्ध हो।

असिद्ध हेतु: ऐसा कारण जो स्वयं ही सिद्ध न हो।

यह वर्गीकरण आधुनिक तार्किक भ्रांतियों (logical fallacies) एवं AI में होने वाले झूठे सहसंबंधों (spurious correlations) से सीधे सम्बद्ध है।

तालिका 1: न्याय दर्शन की अवधारणाओं का AI की समस्याओं एवं समाधानों से तुलनात्मक विश्लेषण

न्याय दर्शन की अवधारणा	AI प्रणाली में समानांतर / प्रकटीकरण	संबद्ध AI नैतिक चुनौती	न्याय-आधारित समाधान का सुझाव
प्रमाण (ज्ञान के स्रोत)			
• प्रत्यक्ष	सेंसर डेटा (इमेज, ऑडियो, टेक्स्ट)।	अपूर्ण/दूषित सेंसर डेटा; प्रतिनिधित्व का अभाव।	**प्रमाण-शुद्धि प्रोटोकॉल:** डेटासेट की विविधता, गुणवत्ता एवं प्रतिनिधित्व का ऑडिट।
• अनुमान	सांख्यिकीय अनुमान (इनफेरेंस) एवं पैटर्न पहचान।	ब्लैक-बॉक्स मॉडल्स; समझ से बाहर तर्क।	**पंचावयवी अनुरूप एक्सप्लेनेशन:** निर्णय के लिए संरचित कारण देना अनिवार्य करना।
• उपमान	सिमिलैरिटी-आधारित मॉडल (जैसे नेयरेस्ट नेबर)।	स्टीरियोटाइपिंग; गलत समानताएँ बिठाना।	**तुलना मानदंडों का स्पष्टीकरण:** 'समानता' को परिभाषित करने वाले मानदंडों का पारदर्शी खुलासा।
• शब्द	ट्रेनिंग डेटासेट एवं प्रोग्रामर द्वारा दिए गए नियम/कोड।	ऐतिहासिक/सामाजिक पूर्वाग्रहों का डेटा में समावेश; दोषपूर्ण प्रोग्रामिंग।	**स्रोत विश्वसनीयता जाँच:** डेटा स्रोतों का नैतिक अनुमोदन; कोड/नियमों की समीक्षा।
पंचावयवी न्याय	एक्सप्लेनेबल AI (XAI) एवं निर्णय लॉग।	अपारदर्शिता; जवाबदेही का अभाव।	**मानकीकृत व्याख्या प्रारूप:** सभी महत्वपूर्ण AI निर्णय पंचावयवी के पाँच चरणों के अनुरूप दर्ज किए जाएँ।
हेत्वाभास (तर्क दोष)	मॉडल की झूठी सहसंबोधन खोजने की प्रवृत्ति।	एल्गोरिदमिक पक्षपात; अतार्किक/हानिकारक निष्कर्ष।	**हेत्वाभास डिटेक्शन यूनिट:** मॉडल टेस्टिंग में ज्ञात तार्किक भ्रांतियों का पता लगाने वाले टेस्ट केस शामिल करना।

कर्तृत्व (एजेंसी)	AI सिस्टम (जो एक नियम-आधारित यंत्र है)।	**दायित्व रिक्तता:** हानि होने पर जिम्मेदारी तय करने में कठिनाई।	**दायित्व आरोपण का सिद्धांत:** AI एक चेतन कर्ता नहीं; अंतिम जवाबदेही मानव कर्ताओं (डेवलपर, डिप्लॉयर, रेगुलेटर) पर होगी।

5. AI की नैतिक चुनौतियों पर न्याय दर्शन का विश्लेषणात्मक अनुप्रयोग

तालिका 1 में दर्शाए गए संबंधों का विस्तार से विश्लेषण यहाँ प्रस्तुत है।

5.1 डेटा पक्षपात एवं प्रमाण की अशुद्धि

AI मॉडल अपने ट्रेनिंग डेटा (जो 'शब्द' प्रमाण का कार्य करता है) में निहित ऐतिहासिक एवं सामाजिक पूर्वाग्रहों को सीखकर सुदृढ़ कर देते हैं। न्याय दर्शन के अनुसार, यदि प्रमाण (यहाँ डेटा) स्वयं दोषपूर्ण है, तो उससे उत्पन्न अनुमान (AI का आउटपुट) कभी भी वैध नहीं हो सकता। एक भर्ती एल्गोरिदम जो ऐतिहासिक रूप से पुरुष-प्रधान उद्योग के डेटा पर प्रशिक्षित है, महिला उम्मीदवारों के प्रति अनजाने में पक्षपात करेगा। न्याय दर्शन इस समस्या के समाधान हेतु "प्रमाण-शुद्धि" पर बल देता है यानि ज्ञान के स्रोत की विश्वसनीयता की पुष्टि करना। AI डेवलपमेंट लाइफसाइकल में इसका अर्थ है डेटासेट की विविधता, प्रतिनिधित्वशीलता एवं गुणवत्ता का गहन ऑडिट किसी 'आप्त' (विश्वसनीय स्रोत) के रूप में इसकी पुष्टि करना), ताकि यह सुनिश्चित हो सके कि 'शब्द' प्रमाण वास्तव में सभी प्रासंगिक समूहों एवं परिदृश्यों का प्रतिनिधित्व करता है।

5.2 पारदर्शिता का अभाव एवं 'ब्लैक बॉक्स' समस्या

जटिल डीप लर्निंग मॉडल्स (जैसे न्यूरल नेटवर्क) के निर्णयन प्रक्रिया को समझ पाना अक्सर कठिन होता है, जिससे विश्वास एवं जवाबदेही का संकट पैदा होता है। न्याय दर्शन की पंचावयवी पद्धति इस समस्या का एक संभावित समाधान प्रस्तुत करती है। यदि प्रत्येक महत्वपूर्ण AI निर्णय को इस पाँच-चरणीय ढाँचे में ढालकर प्रस्तुत किया जाए (एक्सप्लेनेबल AI - XAI का एक मानकीकृत रूप), तो न केवल पारदर्शिता बढ़ेगी, बल्कि मानव ऑपरेटर या रेगुलेटर को तर्क के प्रत्येक चरण विशेषकर 'हेतु' (कारण) और 'उदाहरण' (सार्वभौमिक नियम) की जाँच का अवसर भी मिलेगा।

उदाहरण (एक ऋण आवेदन अस्वीकृति का):

प्रतिज्ञा: "यह ऋण आवेदन अस्वीकृत है।"

हेतु: "क्योंकि आवेदक का क्रेडिट स्कोर 580 है।"

उदाहरण: "सार्वभौमिक नियम/उदाहरण: हमारा मॉडल और ऐतिहासिक डेटा दर्शाता है कि 600 से कम क्रेडिट स्कोर वाले 85% आवेदकों ने ऋण चुकाने में गंभीर चूक की है।"

उपनय: "इस आवेदक का स्कोर 580 है, जो 600 से कम है।"

निगमन: "अतः, यह आवेदन उच्च जोखिम श्रेणी में आता है और अस्वीकृत किया जाता है।"

इस प्रारूप से उपयोगकर्ता यह देख सकता है कि निर्णय केवल एक रहस्यमय संख्या पर आधारित नहीं है, बल्कि एक स्पष्ट कारण और एक (चर्चा के लिए खुले) सामान्य नियम पर आधारित है। अगर 'उदाहरण' (600 का कट-ऑफ) स्वयं पक्षपाती है, तो इस ढाँचे के कारण उस पर सवाल उठाना आसान हो जाता है।

5.3 तर्कसंगत निर्णयन एवं हेत्वाभास से बचाव

AI मॉडल अक्सर झूठे सहसंबंधों को कारण-कार्य सम्बन्ध समझ लेते हैं (जैसे, "हॉस्पिटल में भर्ती होना" और "नमकीन क्रैकर्स का सेवन" दोनों बीमारी के सूचक हैं, पर एक दूसरे का कारण नहीं)। न्याय दर्शन द्वारा वर्गीकृत हेत्वाभास (तार्किक भ्रांतियाँ) इन्हें पहचानने एवं दूर करने का एक व्यवस्थित ढाँचा दे सकती हैं। उदाहरण के लिए, "सव्यभिचार हेतु" वह स्थिति है जहाँ कारण (हेतु) साध्य के साथ अनिवार्य रूप से जुड़ा नहीं है। एक AI मॉडल जो "ज़िप कोड" को "क्रेडिट योग्यता" का कारण मान ले, वह इसी दोष का शिकार है, क्योंकि ज़िप कोड स्वयं क्रेडिट योग्यता का कारण नहीं है, बल्कि यह ऐतिहासिक रेडलाइनिंग जैसे अन्य कारकों से जुड़ा हो सकता है। AI मॉडल्स के ट्रेनिंग एवं वैलिडेशन प्रोटोकॉल में इन भ्रांतियों की पहचान करने वाले टेस्ट केसेस शामिल किए जा सकते हैं।

5.4 जवाबदेही का निर्धारण

जब AI से जुड़ी कोई हानि होती है (जैसे स्वायत्त वाहन दुर्घटना, भर्ती में भेदभाव), तो दायित्व किस पर है? न्याय दर्शन में 'कर्तृत्व' (एजेंसी/कर्ता) की अवधारणा स्पष्ट है केवल चेतन सत्ता ही कर्ता हो सकती है और उसी पर दायित्व होता है। AI एक चेतन सत्ता नहीं है; वह केवल मानव-निर्धारित नियमों (हेतु/एल्गोरिदम) एवं डेटा (प्रमाण) के अनुसार कार्य करता है। अतः, अंतिम दायित्व उन मानव कर्ताओं (डेवलपर्स, डिप्लॉयर्स, रेगुलेटर्स) पर ही रहता है जिन्होंने उन नियमों को चुना, डिज़ाइन किया, अनुमोदित किया या उपयोग की अनुमति दी। यह सिद्धांत AI दायित्व कानून (Liability Law) के लिए एक स्पष्ट दार्शनिक आधार प्रदान करता है: AI स्वयं दोषी नहीं ठहराया जा सकता, परंतु उसे बनाने, बेचने व नियंत्रित करने वाले मानव कर्ता जवाबदेह हैं। पंचावयवी लॉग इस जवाबदेही को ट्रेस करने में सहायक होंगे।

6. प्रस्तावित प्रतिमान: न्याय-आधारित AI ऑडिट एवं डिजाइन ढाँचा

उपरोक्त विश्लेषण के आधार पर, हम एक व्यावहारिक "न्याय-आधारित AI ऑडिट एवं डिजाइन ढाँचा" प्रस्तावित करते हैं। यह ढाँचा AI सिस्टम के जीवनचक्र के प्रत्येक चरण में न्याय दर्शन के सिद्धांतों को समावेशित करने का मार्गदर्शन प्रदान करता है।

AI जीवनचक्र चरण	**न्याय दर्शन के सिद्धांत**	**व्यावहारिक कार्यान्वयन दिशा-निर्देश**	**अपेक्षित परिणाम**
1. आवश्यकता एवं डेटा संग्रह	**प्रमाण-शुद्धि (विशेषकर 'शब्द' प्रमाण)**	• **डेटा स्रोत ऑडिट:** डेटासेट की विविधता, ऐतिहासिक पूर्वाग्रहों की जाँच। • **स्रोत विश्वसनीयता प्रमाणपत्र:** डेटा एकत्रीकरण की नैतिक अनुमोदन प्रक्रिया स्थापित करना।	पक्षपात-रहित, प्रतिनिधि, उच्च-गुणवत्ता वाले प्रशिक्षण डेटासेट।
2. मॉडल डिजाइन एवं प्रशिक्षण	**पंचावयवी संरचना; हेत्वाभास जागरूकता**	• **एक्सप्लेनेशन-बाय-डिजाइन आर्किटेक्चर:** मॉडल में एक मॉड्यूल शामिल करना जो महत्वपूर्ण निर्णयों के लिए पंचावयवी प्रारूप में कारण उत्पन्न करे। • **हेत्वाभास डिटेक्शन टेस्ट सूट:** प्रशिक्षण/वैधीकरण में ज्ञात तार्किक भ्रांतियों व झूठे सहसंबंधों का पता लगाने वाले परिदृश्य शामिल करना।	पारदर्शी तर्क क्षमता वाले मॉडल; गलत अनुमानों की संभावना कम।
3. परीक्षण एवं मूल्यांकन	**निगमन की वैधता की जाँच; 'उदाहरण' का मूल्यांकन**	• **तर्क-मूल्यांकन मेट्रिक्स:** सटीकता के अलावा, मॉडल के तर्क की दृढ़ता, स्थिरता एवं न्यायसंगतता के मात्रात्मक माप। • **न्याय-संगतता परीक्षण:** विभिन्न जनसांख्यिकीय समूहों पर मॉडल के निर्णयों की तुलना करना कि क्या वे एक सार्वभौमिक नियम ('उदाहरण') का पालन करते हैं।	विश्वसनीय, न्यायसंगत एवं तार्किक रूप से ठोस AI मॉडल्स।
4. तैनाती एवं निरंतर निगरानी	**निरंतर प्रमाण-अपडेट; दायित्व ट्रेसबिलिटी**	• **पंचावयवी निर्णय लॉग:** प्रत्येक महत्वपूर्ण निर्णय का संरचित लॉग रखना (प्रतिज्ञा, हेतु, उदाहरण, आदि)।	उच्च जवाबदेही; समय के साथ अनुकूलन एवं सुधार की क्षमता।

		• **प्रतिक्रिया लूप:** वास्तविक दुनिया के नए डेटा ('प्रत्यक्ष' प्रमाण) के आधार पर मॉडल के तर्क ('अनुमान') व नियमों ('उदाहरण') का निरंतर पुनर्मूल्यांकन एवं सुधार।	

इस ढाँचे को अपनाने से AI विकास एक बंद, रहस्यमय प्रक्रिया से एक खुली, विश्लेषण योग्य, और नैतिक रूप से जवाबदेह प्रक्रिया में परिवर्तित हो सकती है। यह डेवलपर्स, रेगुलेटर्स और एंड-यूज़र्स सभी के लिए एक साझा भाषा एवं मूल्यांकन के मानक प्रदान करता है।

7. निहितार्थ, चर्चा एवं सीमाएँ

7.1 निहितार्थ

AI नैतिकता के दार्शनिक आधार का विस्तार: यह शोध AI नैतिकता के वैश्विक विमर्श को एक गहन, गैर-पश्चिमी तार्किक परंपरा से समृद्ध करता है। यह दर्शाता है कि न्याय जैसी अवधारणाओं की व्याख्या केवल पश्चिमी फ्रेमवर्क तक सीमित नहीं है।

व्यावहारिक अभियांत्रिकी समाधानों को दिशा: प्रस्तावित ढाँचा केवल सैद्धांतिक नहीं है। इसके घटक जैसे पंचावयवी एक्सप्लेनेशन जेनरेटर या हेत्वाभास डिटेक्शन टूल्स को तकनीकी रूप से कार्यान्वित किया जा सकता है और मौजूदा एक्सप्लेनेबल AI (XAI) एवं एआई फेयरनेस टूलकिट में एकीकृत किया जा सकता है।

सांस्कृतिक रूप से सूचित नीति निर्माण: यह ढाँचा भारत जैसे देशों के लिए अपनी सांस्कृतिक-दार्शनिक विरासत पर आधारित, स्थानिक संदर्भ के अनुरूप AI नीति व नियम बनाने में एक आधार प्रदान कर सकता है। यह "एक आकार सब पर फिट" (one-size-fits-all) वाले वैश्विक नियमन के लिए एक वैकल्पिक दृष्टिकोण है।

7.2 चर्चा: भारतीय दर्शन के अन्य सिद्धांतों के साथ एकीकरण

न्याय दर्शन का तार्किक ढाँचा अन्य भारतीय दार्शनिक लक्ष्यों के साथ मिलकर कार्य कर सकता है। उदाहरण के लिए:

सांख्य दर्शन का प्रकृति-पुरुष द्वैत- हमें याद दिलाता है कि AI प्रकृति (पदार्थ, एल्गोरिदम) की एक जटिल अभिव्यक्ति है, किंतु इसमें पुरुष (शुद्ध चेतना, आत्मा) नहीं है। इसलिए, इसे चेतन अस्तित्व के समान अधिकार या दर्जा नहीं दिया जा सकता।

वेदान्त का अद्वैत सिद्धांत एवं **बौद्ध दर्शन का अंतर्निर्भरता सिद्धांत**- AI को ऐसे डिजाइन करने की प्रेरणा दे सकता है जो विभाजन के बजाय समग्रता एवं संपूर्ण प्रणाली के कल्याण को बढ़ावा दे।

योग दर्शन का चित्त-नियमन- AI को मानव मन की नकारात्मक वृत्तियों (क्रोध, लालच, भ्रम) को बढ़ाने के बजाय, उन्हें समझने और शांत करने के टूल के रूप में विकसित करने का मार्गदर्शन कर सकता है।

इस प्रकार, न्याय दर्शन तार्किक आधार प्रदान करता है, जबकि अन्य दर्शन नैतिक लक्ष्य एवं दृष्टिकोण देते हैं। दोनों का संयोजन एक समग्र AI दर्शन का निर्माण कर सकता है।

7.3 सीमाएँ

सैद्धांतिक प्रस्ताव: यह एक प्रारंभिक सैद्धांतिक प्रस्ताव है। इस ढाँचे की व्यावहारिक कार्यान्वयन योग्यता, प्रभावकारिता, और संसाधन आवश्यकताओं का परीक्षण भविष्य के प्रायोगिक शोध एवं प्रोटोटाइप विकास के विषय हैं।

अनुवाद की चुनौती: न्याय दर्शन की संस्कृत भाषा में निहित सूक्ष्म अवधारणाओं का समकालीन कम्प्यूटेशनल शब्दावली एवं इंजीनियरिंग प्रथाओं में सटीक अनुवाद एक बौद्धिक चुनौती है।

उद्देश्यों में अंतर: एक दार्शनिक प्रणाली के रूप में न्याय का अंतिम लक्ष्य मोक्ष (व्यक्तिगत मुक्ति) है, जबकि AI नैतिकता का लक्ष्य सामाजिक कल्याण एवं नुकसान से बचाव है। दोनों के उद्देश्यों में यह अंतर सैद्धांतिक समन्वय की आवश्यकता को रेखांकित करता है।

8. निष्कर्ष एवं भविष्य के शोध के क्षेत्र

यह शोध पत्र इस तर्क को स्थापित करता है कि न्याय दर्शन, अपनी तार्किक कठोरता, ज्ञान-मीमांसात्मक स्पष्टता एवं विश्लेषणात्मक पद्धति के कारण, कृत्रिम बुद्धिमत्ता की जटिल नैतिक समस्याओं के समाधान हेतु एक गहन एवं व्यवहार्य दार्शनिक संसाधन है। न्याय दर्शन द्वारा प्रदत्त प्रमाण-सिद्धांत, संरचित तर्क-पद्धति (पंचावयवी), तथा तार्किक भ्रांतियों (हेत्वाभास) का वर्गीकरण, AI प्रणालियों में निहित पक्षपात, अपारदर्शिता एवं अतार्किकता की समस्याओं को मूल स्तर पर यानि दोषपूर्ण ज्ञान प्रक्रियाओं पर हमला करके पहचानने एवं दूर करने का मार्ग प्रशस्त करते हैं। प्रस्तावित 'न्याय-आधारित AI ऑडिट एवं डिजाइन ढाँचा' AI के जीवनचक्र में इस दार्शनिक अंतर्दृष्टि को व्यवहार में लाने की दिशा में एक प्रारंभिक रूपरेखा है।

अंततः, AI का भविष्य केवल तकनीकी उन्नति में नहीं, बल्कि उस नैतिक बुद्धिमत्ता में निहित है जिसके साथ हम उसका निर्माण एवं उपयोग करते हैं। वर्तमान में AI नैतिकता पर वैश्विक विमर्श में गैर-पश्चिमी परंपराओं की उपेक्षा एक गंभीर कमी है । भारतीय चिंतन की समृद्ध न्याय परंपरा इस बौद्धिक एवं नैतिक कमी को दूर करने में एक अमूल्य योगदान दे सकती है। यह हमें याद दिलाती है कि प्रौद्योगिकी का उद्देश्य केवल दक्षता या लाभ नहीं, बल्कि सत्य, न्याय और समग्र कल्याण की प्राप्ति होना चाहिए। भविष्य के शोध एवं नीति-निर्माण में इस दार्शनिक संवाद को आगे बढ़ाने की तत्काल आवश्यकता है।

संदर्भ-सूची

1. Indigenous ethics and artificial intelligence. (2026). AI and Ethics, 6(70). (इस शोध पत्र से उद्धरण: "गैर-पश्चिमी ज्ञान-मीमांसाएँ अभी भी इन बहसों से बहिष्कृत हैं" तथा स्वदेशी नैतिकता पर चर्चा)।
2. Singler, B., & Watts, F. (2024). Hinduism and Artificial Intelligence. In The Cambridge Companion to Religion and Artificial Intelligence. Cambridge University Press.
3. Upadhyay, R. K., Upadhyay, A., Upadhay, D. K., Saxena, M., & Sonkar, S. S. (n.d.). Artificial Intelligence Ethics: Navigating the Moral Challenges of AI. Propulsion Tech Journal.
4. Ramiz, R. (2025, Oct 8). Artificial Intelligence and Ethics: Moral Challenges from a Philosophical Perspective. Philosophy Study.
5. Shah, K., Joshi, H., & Joshi, H. (2025). Integrating Moral Values in AI: Addressing Ethical Challenges for Fair and Responsible Technology. Journal of Informatics and Web Engineering, 4(1), 213–227.
6. Gautama. Nyaya Sutra. (मूल संस्कृत ग्रंथ).
7. Vatsyayana. Nyaya Bhasya. (मूल संस्कृत भाष्य).
8. Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In The Cambridge Handbook of Artificial Intelligence (pp. 316-334). Cambridge University Press.
9. Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review.
10. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).